ELSEVIER

Fuzzy Sets and Systems 123 (2001) 343–358

FUZZY sets and systems

www.elsevier.com/locate/fss

# Data-based fuzzy rule test for fuzzy modelling

Angelika Krone[a,*], Heike Taeger[b]

[a]Faculty of Electrical Engineering, Department of Control Engineering, Prof. Kiendl (ESR), University of Dortmund, 44221 Dortmund, Germany

[b]Faculty of Statistics, University of Dortmund, Germany

Received 6 July 1999; received in revised form 5 June 2000; accepted 23 June 2000

**Abstract**

In the field of fuzzy modelling, the exclusive consideration of the modelling error leads to problems concerning the handling of high-dimensional applications and the interpretability of the resulting rule base. To solve those problems, a statistically motivated fuzzy rule test is proposed. It decides if a fuzzy IF/THEN statement is a relevant rule or not. In this way, the problem of finding a good rule base can be reduced to the problem of finding good, relevant rules. 

*Keywords:* Fuzzy statistics and data analysis; Fuzzy system models; Learning; Confidence intervals for fuzzy events

## 1. Introduction

Applying data-based fuzzy modelling methods to industrial applications, the following two points must be considered:

- There are often many possible input variables resulting in enormous search spaces and rule bases.
- On the part of industrial operators, interpretable results are desired that allow insight.

A fuzzy rule test, which decides on the basis of the available learning data whether a fuzzy IF/THEN statement represents a locally important aspect of the dependency between the input and output variables, can help on both accounts:

- A fuzzy rule test allows the enormous problem of finding a complete fuzzy rule base to be broken down to the much smaller problem of finding single fuzzy rules for incremental collection (Fig. 1) [14,18]. The search space of single fuzzy rules is significantly smaller than the search space of complete fuzzy rule bases, and also increases at a significantly lower rate with the number of input variables. Thus, many more input variables can be handled in a given amount of computing time. Furthermore, a fuzzy rule test can

* Corresponding author. Tel.: +49-231-755-3762; fax: +49-231-755-2752.
*E-mail address:* winrosa@esr.e-technik.uni-dortmund.de (A. Krone).

0165-0114/01/$ - see front matter 
PII: S0165-0114(00)00112-3

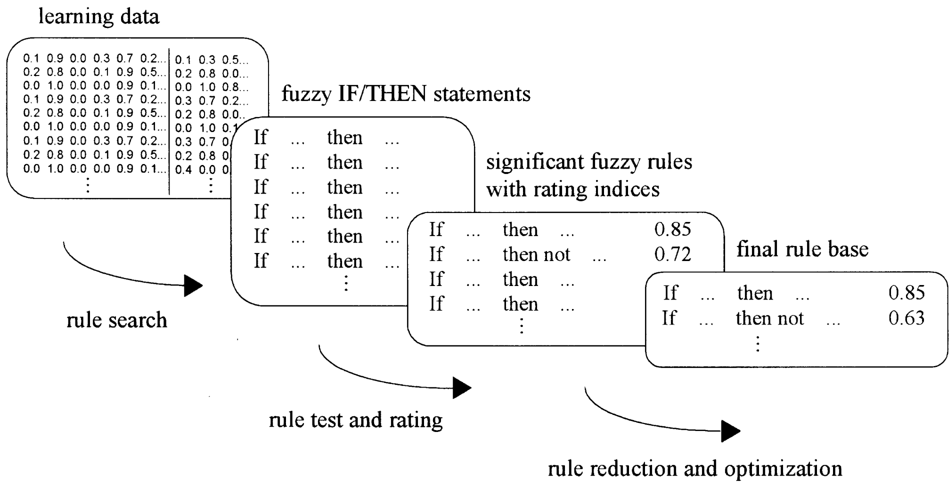



Fig. 1. Embedding of the rule test in the fuzzy modelling process.

determine whether several input situations can be covered by one fuzzy rule instead of several distinct rules, allowing the final fuzzy rule base to be made as small as possible [11].

- A fuzzy rule test allows a fuzzy rule base to be built exclusively out of locally reasonable rules that can be interpreted by the industrial operators and responsible managers. The interpretability encourages the acceptance of new components and facilitates adjustment to change. Many fuzzy modelling methods presented in the literature consider only the input/output behaviour, which often results in fuzzy rules that are not locally reasonable.

Depending on the field of application, a fuzzy rule test can have different intentions, for example:

- It is intended that a fuzzy rule represents a relevant dependency between the input situation in the premise and the output in the conclusion for the purpose of a causal relation.
- It is intended that a fuzzy rule has a good hit rate.
- It is intended that a fuzzy rule predicts the mean value.

Normally, those intentions are inconsistent. For example, the rule ‘If Peter does not eat up, it will not be sunny tomorrow’ will have a good hit rate of about 80% if the conclusion is true, irrespectively of the premise in 80% of the days. However, it does not represent a causal relation.

Here, a fuzzy rule test is developed that pursues the first point, that is, finding relevant rules for the purpose of causal relations. As its task is to separate the relevant from irrelevant fuzzy rules, it is called a fuzzy relevance test. In order to handle contradictory learning data, the aim is to profit from statistical methods.

A relevance test for crisp rules, based on the computation of confidence intervals, has been introduced by Kiendl and Krabs [10]. This concept is presented in Section 2. For fuzzy modelling, the formula of the crisp relevance test can be algorithmically extended to the use of fuzzy values (Section 3). This approach has the advantage that it is immediately available and the computing time is not higher as in the crisp case. However, the statistical verification is no longer applicable, and it is questionable how the results can be interpreted. Thus, a fuzzy approach has been developed (Section 4) in which confidence intervals are computed by two different methods. The results of the algorithmic extension and the two fuzzy approaches are illustrated and compared (Section 5). As a conclusion, each approach is assigned a special field of application.

## 2. Crisp relevance test

A statement of the following form is to be examined:

IF $S$ THEN $C$

$S$ represents an input situation, $C$ an output event. The input situation, resp. the output event is true or not true. The corresponding characteristic functions $\mathbf{I}_S$ and $\mathbf{I}_C$ take the value 1 if the input situation, resp. the output event is true and the value 0 if the input situation, resp. the output event is not true:

$$\mathbf{I}_S(X(k)) = \begin{cases} 1\text{:} & S \text{ is true for the data sample } d_k, \\ 0\text{:} & S \text{ is not true for the data sample } d_k, \end{cases}$$

$$\mathbf{I}_C(Y(k)) = \begin{cases} 1\text{:} & C \text{ is true for the data sample } d_k, \\ 0\text{:} & C \text{ is not true for the data sample } d_k. \end{cases}$$

$X=(X_1,X_2,\dots)$ is the vector with the input variables. $Y$ is the output variable. The premise and the conclusion refer to the same data sample $d_k=(x_1(k),x_2(k),\dots,y(k))=(x(k),y(k))$. It includes the realizations of $X(k)$, $Y(k)$ belonging together, for example the observations at the date $k$.

**Example.** $X_1$ is the heating temperature, $X_2$ is the outdoor temperature, $Y$ is the room temperature. There are five data samples $(n=5)$ with $d_1=(50^\circ\text{C},-10^\circ\text{C},14^\circ\text{C})$, $d_2=(35^\circ\text{C},-5^\circ\text{C},12^\circ\text{C})$, $d_3=(25^\circ\text{C},-2^\circ\text{C},8^\circ\text{C})$, $d_4=(20^\circ\text{C},-5^\circ\text{C},5^\circ\text{C})$, $d_5=(25^\circ\text{C},-10^\circ\text{C},3^\circ\text{C})$. A possible IF/THEN statement is: IF ((heating temperature is lower than 30°C) $\wedge$ (outdoor temperature is lower than 0°C)) THEN (room temperature is under 10°C). This corresponds to: IF $(X_1<30^\circ\text{C}\wedge X_2<0^\circ\text{C})$ THEN $(Y<10^\circ\text{C})$. The characteristic functions are defined by

$$\mathbf{I}_S(X(k)) = \begin{cases} 1\text{:} & X_1(k) < 30^\circ\text{C} \wedge X_2(k) < 0^\circ\text{C}, \\ 0\text{:} & \text{else} \end{cases} \qquad \mathbf{I}_C(Y(k)) = \begin{cases} 1\text{:} & Y(k) < 10^\circ\text{C}, \\ 0\text{:} & \text{else.} \end{cases}$$

For the five data samples, the following values result: $\mathbf{I}_S(x(1))=0$, $\mathbf{I}_S(x(2))=0$, $\mathbf{I}_S(x(3))=1$, $\mathbf{I}_S(x(4))=1$, $\mathbf{I}_S(x(5))=1$, $\mathbf{I}_C(y(1))=0$, $\mathbf{I}_C(y(2))=0$, $\mathbf{I}_C(y(3))=1$, $\mathbf{I}_C(y(4))=1$, $\mathbf{I}_C(y(5))=1$.

The probability that the output event is true $(\mathbf{I}_C(Y(k))=1)$ is $p=P(C)$. The conditional probability that the output event $C$ is true under the condition that the input situation $S$ is true $(\mathbf{I}_C(Y(k))=1|\mathbf{I}_S(X(k))=1)$ is $p_\lambda=P(C|S)$. The more these two probabilities differ the more the IF/THEN statement can be seen as relevant.

As these probabilities are not known, they are estimated on the basis of the data samples $d_k$ by the relative frequencies:

$$\hat{p}=\frac{m}{n} \quad \text{and} \quad \hat{p}_\lambda=\frac{m_\lambda}{n_\lambda}$$

with

$$n := \text{number of data samples } d_k,$$

$$m := \sum_{k=1}^{n} \mathbf{I}_C(Y(k)),$$

$$n_\lambda := \sum_{k=1}^{n} \mathbf{I}_S(X(k)),$$

$$m_\lambda := \sum_{k=1}^{n} (\mathbf{I}_S(X(k)) \wedge \mathbf{I}_C(Y(k))).$$

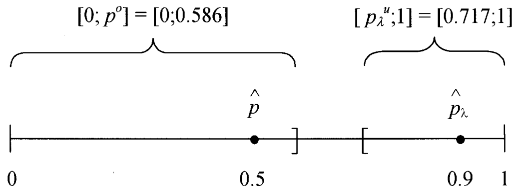



Fig. 2. Confidence limits for $n$=100, $m$=50, $n_\lambda = 20$ and $m_\lambda = 18$ calculated by the crisp relevance test.

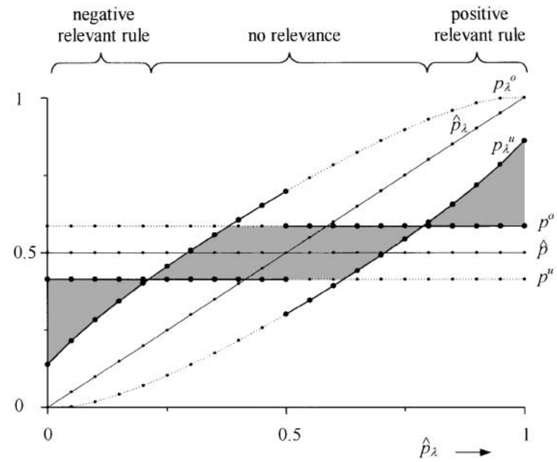



Fig. 3. Confidence limits for $n = 100$, $m = 50$, $n_\lambda = 20$ and $m_\lambda = 0, 1, 2, \ldots, 20$ calculated by the crisp relevance test.

For $X(1),\ldots,X(n)$ independent, identically distributed (i.i.d.) and $Y(1),\ldots,Y(n)$ i.i.d., it can be proved that $\hat{p}$ and $\hat{p}_\lambda$ are consistent and uniformly minimal-variance unbiased estimators.

$\mathbf{I}_S(X)$ resp. $\mathbf{I}_C(Y)$ are Bernoulli distributed with the parameter $p_\lambda$ resp. $p$. On this basis, confidence intervals can be calculated for $p$ and $p_\lambda$ with the Pearson–Clopper values [7]. They cover the probabilities $p$ and $p_\lambda$ each with a given probability $1 - \alpha$ (confidence coefficient).

As only one side of the confidence intervals is interesting in each relevance test, one or the other of the one-sided confidence intervals $I^{\mathrm{o}} := [0; p^{\mathrm{o}}]$, $I_\lambda^{\mathrm{u}} := [p_\lambda^{\mathrm{u}}; 1]$ or $I_\lambda^{\mathrm{o}} := [0; p_\lambda^{\mathrm{o}}]$, $I^{\mathrm{u}} := [p^{\mathrm{u}}; 1]$ is calculated.

In the case,

$$\hat{p} < \hat{p}_\lambda \wedge p^{\mathrm{o}} < p_\lambda^{\mathrm{u}}$$

the statement 'IF $S$ THEN $C$' is a *positive relevant rule*. In the case

$$\hat{p} > \hat{p}_\lambda \wedge p^{\mathrm{u}} > p_\lambda^{\mathrm{o}},$$

the negative statement 'IF $S$ THEN $\neg C$' is a *negative relevant rule* [9]. In all other cases, it is true that

$$[p^{\mathrm{u}}; p^{\mathrm{o}}] \cap [p_\lambda^{\mathrm{u}}; p_\lambda^{\mathrm{o}}] \neq \emptyset$$

and thus no relevant rule can be extracted.

After the relevance test, the relevant rules can be assigned a rating index [13,14].

**Example.** There are 100 data samples ($n = 100$). The output event $C$ is true in 50 of the 100 data samples ($m = 50$), so that $\hat{p} = 0.5$. The input situation $S$ is true in 20 of the 100 data samples ($n_\lambda = 20$). In 18 of the 20 data samples the output event $C$ is true ($m_\lambda = 18$), so that $\hat{p}_\lambda = 0.9$. As $\hat{p} < \hat{p}_\lambda$, the interval limits $p^{\mathrm{o}}$ and $p_\lambda^{\mathrm{u}}$ must be calculated. With a confidence coefficient of 0.95, one gets $p^{\mathrm{o}} = 0.586$ and $p_\lambda^{\mathrm{u}} = 0.717$. The result is visualized in Fig. 2. The statement 'IF $S$ THEN $C$' is a positive relevant rule as the confidence intervals do not intersect. The results for all possible values of $m_\lambda$ $(0, 1, 2, \ldots, 20)$ are visualized in Fig. 3.

## 3. Algorithmic extension of the crisp relevance test

If the input situation and the output event are described by fuzzy sets [22], they can either be true or not true, but also can be true to a certain degree, normally between 0 and 1. Thus, the characteristic functions $\mathbf{I}_S(X(k))$ and $\mathbf{I}_C(Y(k))$ are substituted by the membership functions:

$$\mu_S(X(k)) \in [0;1] \quad \text{and} \quad \mu_C(Y(k)) \in [0;1].$$

In the case of fuzzy input situations and fuzzy output events, the formulae of the crisp relevance test can be extended algorithmically from integer to real values. This extension is statistically not justified as the $\mu_S(X)$ resp. $\mu_C(Y)$ are no longer Bernoulli distributed. Nevertheless, one achieves a type of interpolating solution that calculates the correct statistical values in the special case of crisp fuzzy sets (characteristic functions).

Then, the estimators are given by

$$\hat{p} = \frac{m}{n} \quad \text{and} \quad \hat{p}_\lambda = \frac{m_\lambda}{n_\lambda}$$

with

$$n := \text{number of data samples } d_k,$$

$$m := \sum_{k=1}^{n} \mu_C(Y(k)),$$

$$n_\lambda := \sum_{k=1}^{n} \mu_S(X(k)),$$

$$m_\lambda := \sum_{k=1}^{n} (\mu_S(X(k)) \wedge \mu_C(Y(k))).$$

The '$\wedge$' operator can be realized by one of the numerous fuzzy AND operators [16]. Most reasonably, it should be the one that is also used to calculate $\mu_S(x(k))$ from the individual degrees of activation of the different input values $x_i(k)$.

The real values $m, n_\lambda$ and $m_\lambda$ are inserted in the formulae of the crisp calculation of the confidence intervals, though the formulae are only defined for integer values of $m, n_\lambda$ and $m_\lambda$. As an example, Fig. 4 shows the resulting interpolation between the crisp values for the interval limit $p_\lambda^{\mathrm{u}}$.

## 4. Fuzzy relevance test

It can immediately be questioned whether the above extension of the crisp algorithm actually gives meaningful results. Therefore, a relevance test for fuzzy rules has been developed, using a statistical approach. It examines rules of the form

IF $S$ THEN $C$

where $S$ represents a fuzzy input situation described by the membership function $\mu_S(X)$ and $C$ a fuzzy output event described by the membership function $\mu_C(Y)$.

In accordance with the methodology of the crisp relevance test, adequate probabilities and estimators must be defined first (Section 4.1). Afterwards, a method for calculating the confidence intervals must be developed. In contrast to the crisp case, as the distributions of $\mu_S(X)$ resp. $\mu_C(Y)$ are not known, an exact

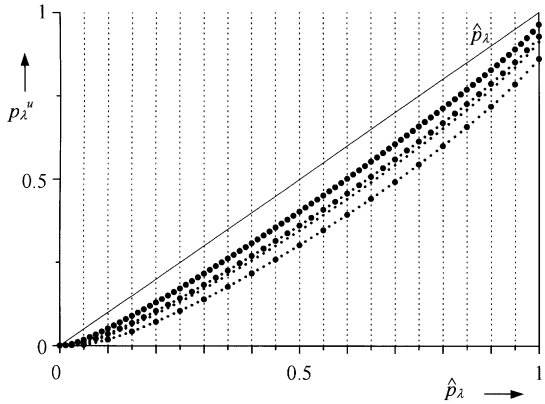



Fig. 4. Interpolating values of the algorithmic extension (represented by small circles '·') and crisp values (represented by big circles '•'), examplarily shown for the interval limit $p_\lambda^u$: • ··· • $n_\lambda = 20$ ($m_\lambda = 0, 0.25, 0.5, 0.75, 1, 1.25, \ldots, 20$), ··· $n_\lambda = 32.5$ ($m_\lambda = 0, \frac{32.5}{80}, \frac{2\cdot 32.5}{80}, \frac{3\cdot 32.5}{80}, \ldots, 32.5$), • · • $n_\lambda = 40$ ($m_\lambda = 0, 0.5, 1, 1.5, 2, \ldots, 40$), • • • $n_\lambda = 80$ ($m_\lambda = 0, 1, 2, 3, \ldots, 80$).

parametric calculation of the confidence intervals is not possible. Nevertheless, two different approaches can be made [21]:

- a non-parametric calculation,
- an asymptotic calculation.

In Section 4.2, the first approach is pursued by using a Bootstrap method for the calculation of the confidence intervals. In Section 4.3, the second approach is pursued by using the central limit theorem and the Fieller method.

### 4.1. Probabilities and estimators for fuzzy events

Zadeh [23] defines the probability of a fuzzy event $A$ by

$$P(A) = \int_A f(z)\,\mathrm{d}z = \int_{\mathbf{R}} \mu_A(z) f(z)\,\mathrm{d}z = E[\mu_A(Z)],$$

where $Z$ is the random variable, $f(z)$ the density of $Z$, $\mu_A(Z)$ the membership function for the fuzzy event $A$, and $E[\cdot]$ the expected value.

Other authors have been keen to take this suggestion [2,20]. In [20] it is proved that the Kolmogoroff axioms of a probability are fulfilled for finite event spaces.

On this basis, the probability of the fuzzy output event $C$ is

$$P(C) = E[\mu_C(Y)].$$

The conditional probability of the fuzzy output event $C$ under the fuzzy situation $S$ is

$$P(C|S) = \frac{P(C \cap S)}{P(S)} = \frac{E[\mu_{C\cap S}(Y,X)]}{E[\mu_S(X)]} = \frac{E[\mu_C(Y) \wedge \mu_S(X)]}{E[\mu_S(X)]} \quad \text{with } P(S) \neq 0.$$

As the '$\wedge$' operator, only the algebraic product

$$\mu_C(Y) \wedge \mu_S(X) = \mu_C(Y)\mu_S(X)$$

makes sense in the field of probabilities, as it is the only operator that can fulfil the following two statistical equations [1]:

1. $P(C\cap S)+P(\bar{C}\cap S)=P(S)$,
2. $P(C\cap S)=P(C)P(S)$ if $C$ and $S$ are independent fuzzy events

The complement is defined by $\mu_{\bar{C}}(Y)=1-\mu_C(Y)$.

An estimator for the probability $P(C)$ is [1,2]

$$\hat{p}=\frac{1}{n}\sum_{k=1}^{n}\mu_C(Y(k))=\frac{m}{n}.$$

An estimator for the probability $P(C|S)$ is

$$\hat{p}_\lambda=\frac{\sum_{k=1}^{n}(\mu_C(Y(k))\mu_S(X(k)))}{\sum_{k=1}^{n}\mu_S(X(k))}=\frac{m_\lambda}{n_\lambda}.$$

For $\mu_C(Y(1)),\ldots,\mu_C(Y(n))$ i.i.d. and $\mu_S(X(1)),\ldots,\mu_S(X(n))$ i.i.d. it can be proved that $\hat{p}$, $m_\lambda$ and $n_\lambda$ are consistent and unbiased estimators [21]. They can be interpreted as average degrees of membership.

Comparing these estimators with those of the algorithmic generalization of the crisp relevance test for fuzzy values; it can be seen that the formulae of the estimators are identical, if the algebraic product is chosen as the '$\wedge$' operator.

### 4.2. Bootstrap fuzzy relevance test

The Bootstrap methods are resampling methods suggested by Efron [3–5]. Among other applications, they can serve to calculate confidence intervals. The name Bootstrap is derived from one of the tales of Baron von Münchhausen, who is said to have pulled himself out of a swamp by his bootstraps. For the relevance test, the non-parametric Bootstrap method $\mathrm{BC_a}$ (bias-corrected and accelerated) [4,8] is used.

From the $n$ data samples $d_k=(x_1(k),x_2(k),\ldots,y(k))$ represented by $(d_1,d_2,\ldots,d_k,\ldots,d_n)$, $w$ random samples of the size $n$ (called Bootstrap samples) are drawn with replacement:

$$\begin{aligned}&(d_1^{*(1)},d_2^{*(1)},\ldots,d_n^{*(1)}),\\&(d_1^{*(2)},d_2^{*(2)},\ldots,d_n^{*(2)}),\\&\qquad\vdots\\&(d_1^{*(w)},d_2^{*(w)},\ldots,d_n^{*(w)}).\end{aligned}$$

For each Bootstrap sample, the estimators for $P(C)$ and $P(C|S)$ are calculated:

$$\begin{aligned}&\hat{p}^{*(1)},\hat{p}_\lambda^{*(1)},\\&\hat{p}^{*(2)},\hat{p}_\lambda^{*(2)},\\&\qquad\vdots\\&\hat{p}^{*(w)},\hat{p}_\lambda^{*(w)}.\end{aligned}$$

The Bootstrap replications $\hat{p}^{*(1)},\ldots,\hat{p}^{*(w)}$ and $\hat{p}_\lambda^{*(1)},\ldots,\hat{p}_\lambda^{*(w)}$ are sorted in ascending order. Then, the limits of the one-sided confidence intervals are the following:

$$p^{\mathrm{u}} = \hat{p}^{*(g_{\mathrm{u}})} \quad (= \text{the } g_{\mathrm{u}}\text{th smallest value of } \hat{p}^{*(1)},\ldots,\hat{p}^{*(w)}),$$
$$p^{\mathrm{o}} = \hat{p}^{*(g_{\mathrm{o}})} \quad (= \text{the } g_{\mathrm{o}}\text{th smallest value of } \hat{p}^{*(1)},\ldots,\hat{p}^{*(w)}),$$
$$p_\lambda^{\mathrm{u}} = \hat{p}_\lambda^{*(g_{\lambda\mathrm{u}})} \quad (= \text{the } g_{\lambda\mathrm{u}}\text{th smallest value of } \hat{p}_\lambda^{*(1)},\ldots,\hat{p}_\lambda^{*(w)}),$$
$$p_\lambda^{\mathrm{o}} = \hat{p}_\lambda^{*(g_{\lambda\mathrm{o}})} \quad (= \text{the } g_{\lambda\mathrm{o}}\text{th smallest value of } \hat{p}_\lambda^{*(1)},\ldots,\hat{p}_\lambda^{*(w)})$$

with

$$g_{\mathrm{u}} := \mathrm{trunc}(\beta_{\mathrm{u}}(w+1)),$$
$$g_{\mathrm{o}} := \mathrm{trunc}(\beta_{\mathrm{o}}(w+1)),$$
$$g_{\lambda\mathrm{u}} := \mathrm{trunc}(\beta_{\lambda\mathrm{u}}(w+1)),$$
$$g_{\lambda\mathrm{o}} := \mathrm{trunc}(\beta_{\lambda\mathrm{o}}(w+1)),$$
$$\mathrm{trunc}(v) := \text{whole-numbered part of } v.$$

The $\beta$ values are calculated by the distribution function $\Phi$ of the standard normal distribution:

$$\beta_{\mathrm{u}} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1-\hat{a}(\hat{z}_0 + z^{(\alpha)})}\right),$$

$$\beta_{\mathrm{o}} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right),$$

$$\beta_{\lambda\mathrm{u}} = \Phi\left(\hat{z}_{\lambda 0} + \frac{\hat{z}_{\lambda 0} + z^{(\alpha)}}{1-\hat{a}_\lambda(\hat{z}_{\lambda 0} + z^{(\alpha)})}\right),$$

$$\beta_{\lambda\mathrm{o}} = \Phi\left(\hat{z}_{\lambda 0} + \frac{\hat{z}_{\lambda 0} + z^{(1-\alpha)}}{1-\hat{a}_\lambda(\hat{z}_{\lambda 0} + z^{(1-\alpha)})}\right),$$

where $z^{(\alpha)}$ and $z^{(1-\alpha)}$ are, respectively, the $\alpha$-, $(1-\alpha)$-quantile of the standard normal distribution; $1-\alpha$ is the confidence coefficient; $\hat{z}_0$ and $\hat{z}_{\lambda 0}$ are the bias parameters; and $\hat{a}$ and $\hat{a}_\lambda$ are the acceleration parameters.

The bias parameters are calculated by the following quantiles of the standard normal distribution

$$\hat{z}_0 = z^{\left(\frac{r}{w}\right)},$$

$$\hat{z}_{\lambda 0} = z^{\left(\frac{r_\lambda}{w}\right)}$$

where $r$ is the number of Bootstrap replications $\hat{p}^{*(\cdot)}$ that are lower than $\hat{p}$, and $r_\lambda$ the number of Bootstrap replications $\hat{p}_\lambda^{*(\cdot)}$ that are lower than $\hat{p}_\lambda$.

The acceleration parameters are calculated by

$$\hat{a} = \frac{\sum_{l=1}^{n}(\hat{p}^{(-)} - \hat{p}^{(l)})^3}{6[\sum_{l=1}^{n}(\hat{p}^{(-)} - \hat{p}^{(l)})^2]^{3/2}},$$

$$\hat{a}_\lambda = \frac{\sum_{l=1}^{n}(\hat{p}_\lambda^{(-)} - \hat{p}_\lambda^{(l)})^3}{6[\sum_{l=1}^{n}(\hat{p}_\lambda^{(-)} - \hat{p}_\lambda^{(l)})^2]^{3/2}}$$

where $\hat{p}^{(l)}$ and $\hat{p}_{\lambda}^{(l)}$ are the estimators on the basis of the $l$th Bootstrap sample $(d_1,\ldots,d_{l-1},d_{l+1},\ldots,d_n)$, $\hat{p}^{(-)} := \frac{1}{n}\sum_{l=1}^{n}\hat{p}^{(l)}$, and $\hat{p}_{\lambda}^{(-)} := \frac{1}{n}\sum_{l=1}^{n}\hat{p}_{\lambda}^{(l)}$.

The $BC_a$ confidence intervals are second-order accurate [17].

In this context, a fundamental disadvantage of the Bootstrap method is the necessary computing time, as a minimum number of Bootstrap samples for the calculation of confidence intervals is $w = 1000$ [5]. Consequently, for high values of $n$ and a high number of IF/THEN statements the method is not practicable.

Diagrams such as Fig. 3 are not possible for the Bootstrap method as the results are dependent on concrete data samples. Results are calculated for an example in Section 5.

### 4.3. Asymptotic fuzzy relevance test

The conventional distribution functions are not adequate for $\mu_S(X)$ and $\mu_C(Y)$. The beta distribution comes nearest, as it has values between 0 and 1. However, it ignores that $\mu_S(X)$ and $\mu_C(Y)$ are partly discretely (0, 1) and partly continuously (]0; 1[) distributed. Nevertheless, one could calculate confidence intervals for $E[\mu_C(Y)]$, which is distributed according to the sum of beta distributed variables. However, the resulting distribution of the quotient $E[\mu_C(Y)\mu_S(X)]/E[\mu_S(X)]$ cannot be derived easily, so that confidence intervals for the conditional probability cannot be calculated.

Another possibility is to assume forthwith a distribution for $E[\mu_C(Y)]$ instead of for $\mu_C(Y)$. According to the central limit theorem [7], the distribution of the sum of any distributed variables converges to a normal distribution for $n$ converging to infinity, so it can be shown that the following is valid:

$$\frac{\frac{1}{n}\sum_{k=1}^{n}(\mu_C(Y(k)) - E[\mu_C(Y(k))])}{\sqrt{VAR[\mu_C(Y(k))]}}\sqrt{n} \overset{n\to\infty}{\sim} \mathrm{N}(0,1)$$

where $n$ is the number of data samples $d_k$, $VAR[\cdot]$ the variance, and $\mathrm{N}(0,1)$ the standard normal distribution.

An approximation to a normal distribution can already be obtained for smaller values of $n$. Experiments with different process data have shown that the approximation is sufficiently close for more than 40 data samples.

As a conclusion, $p^{\mathrm{u}}$ and $p^{\mathrm{o}}$ can be calculated asymptotically for $E[\mu_C(Y)]$ for $Y(1),\ldots,Y(n)$ i.i.d. by

$$p^{\mathrm{u}} = \max\left\{0, \frac{m}{n} - \frac{t_{n-1;1-\alpha}}{\sqrt{n}} S_C\right\},$$

$$p^{\mathrm{o}} = \min\left\{\frac{m}{n} + \frac{t_{n-1;1-\alpha}}{\sqrt{n}} S_C, 1\right\}$$

with

$n :=$ number of data samples $d_k$,

$t_{n-1;1-\alpha} := (1-\alpha)$ quantile of the $t$ distribution with $(n-1)$ degrees of freedom,

$1-\alpha :=$ confidence coefficient,

$$S_C := \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(\mu_C(Y(k) - \frac{m}{n}\right)^2} \text{ (estimator for standard deviation)},$$

$$m := \sum_{k=1}^{n}\mu_C(Y(k)).$$

For the conditional probability, the calculation is more difficult because of the quotient. According to the central limit theorem, $m_\lambda$ and $n_\lambda$ are asymptotically and normally distributed. The Fieller method [6] can then be applied for $X(1),\dots,X(n)$ i.i.d. and $Y(1),\dots,Y(n)$ i.i.d.

The following asymptotic confidence limits result from this approach:

$$p_\lambda^{\mathrm{u}} = \max\left\{0, \frac{\dfrac{m_\lambda n_\lambda}{n} - t^2_{n-1;1-\alpha} S_{CS,S}}{\dfrac{n_\lambda^2}{n} - t^2_{n-1;1-\alpha} S_S^2} - \frac{\sqrt{\left(\dfrac{m_\lambda n_\lambda}{n^2} - \dfrac{t^2_{n-1;1-\alpha}}{n} S_{CS,S}\right)^2 - \left(\left(\dfrac{m_\lambda}{n}\right)^2 - \dfrac{t^2_{n-1;1-\alpha}}{n} S_{CS}^2\right)\left(\left(\dfrac{n_\lambda}{n}\right)^2 - \dfrac{t^2_{n-1;1-\alpha}}{n} S_S^2\right)}}{\left(\dfrac{n_\lambda}{n}\right)^2 - \dfrac{t^2_{n-1;1-\alpha}}{n} S_S^2}\right\}$$

$$p_\lambda^{\mathrm{o}} := \min\left\{\frac{\dfrac{m_\lambda n_\lambda}{n} - t^2_{n-1;1-\alpha} S_{CS,S}}{\dfrac{n_\lambda^2}{n} - t^2_{n-1;1-\alpha} S_S^2} + \frac{\sqrt{\left(\dfrac{m_\lambda n_\lambda}{n^2} - \dfrac{t^2_{n-1;1-\alpha}}{n} S_{CS,S}\right)^2 - \left(\left(\dfrac{m_\lambda}{n}\right)^2 - \dfrac{t^2_{n-1;1-\alpha}}{n} S_{CS}^2\right)\left(\left(\dfrac{n_\lambda}{n}\right)^2 - \dfrac{t^2_{n-1;1-\alpha}}{n} S_S^2\right)}}{\left(\dfrac{n_\lambda}{n}\right)^2 - \dfrac{t^2_{n-1;1-\alpha}}{n} S_S^2}, 1\right\}$$

with

$n :=$ number of data samples $d_k$,

$$m_\lambda := \sum_{k=1}^{n} \mu_C(Y(k))\mu_S(X(k)),$$

$$n_\lambda := \sum_{k=1}^{n} \mu_S(X(k)),$$

$$S_{CS}^2 := \frac{1}{n-1} \sum_{k=1}^{n} \left(\mu_C(Y(k))\mu_S(X(k)) - \frac{m_\lambda}{n}\right)^2,$$

$$S_S^2 := \frac{1}{n-1} \sum_{k=1}^{n} \left(\mu_S(X(k)) - \frac{n_\lambda}{n}\right)^2,$$

$$S_{CS,S} := \frac{1}{n-1} \sum_{k=1}^{n} \left(\mu_C(Y(k))\mu_S(X(k)) - \frac{m_\lambda}{n}\right)\left(\mu_S(Y(k)) - \frac{n_\lambda}{n}\right),$$

$t_{n-1;1-\alpha} := (1-\alpha)$ quantile of the $t$-distribution with $(n-1)$ degrees of freedom,

$1-\alpha :=$ confidence coefficient.

It can be shown that the result for the unconditional probability is a special case of the result for the conditional probability with $S=\Omega$ and $\mu_\Omega = 1$ [21].

A main difference from the crisp relevance test is that the quantities $m, n, m_\lambda, n_\lambda$ are not sufficient to calculate the confidence intervals. The estimated variances $S_C^2, S_S^2, S_{CS}^2, S_{CS,S}$ are also necessary. Thus, for one combination of $m, n, m_\lambda, n_\lambda$ an infinite number of values for the confidence limits is possible.

For the unconditional probability, the smallest confidence intervals are achieved for $S_C^2 = 0$. Then, the confidence limits are $p^u = p^o = \hat{p}$. The largest confidence intervals are achieved if $\mu_C(Y(k)) \in \{0;1\}$ and $\mu_S(X(k)) \in \{0;1\}$ are valid for all values of $k$. Then, the variance $S_C^2$ becomes maximum. So, the range of values for the confidence limits of $p$ is given by

$$\max\left\{0, \frac{m}{n} - \frac{t_{n-1;1-\alpha}}{\sqrt{n(n-1)}}\sqrt{m - \frac{m^2}{n}}\right\} \leqslant p^u \leqslant \frac{m}{n}$$

$$\frac{m}{n} \leqslant p^o \leqslant \min\left\{\frac{m}{n} + \frac{t_{n-1;1-\alpha}}{\sqrt{n(n-1)}}\sqrt{m - \frac{m^2}{n}}, 1\right\}.$$

For the conditional probability, the smallest confidence intervals are achieved for $S_S^2 = 0$, $S_{CS}^2 = 0$, $S_{CS,S} = 0$. Then, the confidence limits are $p_\lambda^u = p_\lambda^o = \hat{p}_\lambda$. Analyses show that the largest confidence intervals are achieved if $\mu_C(Y(k)) \in \{0;1\}$ and $\mu_S(X(k)) \in \{0;1\}$ is valid for all values of $k$. Then, the variances $S_S^2$, $S_{CS}^2$, $S_{CS,S}$ become maximum. A proof has not yet been obtained. However, assuming the correctness of that relationship, the range of values for the confidence limits of $p_\lambda$ is given by

$$\max\left\{0, \frac{m_\lambda}{n_\lambda} - \sqrt{\frac{m_\lambda^2}{n_\lambda^2} - \frac{m_\lambda^2\left[\left(1 - \frac{t_{n-1;1-\alpha}^2}{m_\lambda}\right)n - 1 + t_{n-1;1-\alpha}^2\right]}{n_\lambda^2\left[\left(1 - \frac{t_{n-1;1-\alpha}^2}{n_\lambda}\right)n - 1 + t_{n-1;1-\alpha}^2\right]}}\right\} \leqslant p_\lambda^u \leqslant \frac{m_\lambda}{n_\lambda}$$

$$\frac{m_\lambda}{n_\lambda} \leqslant p_\lambda^o \leqslant \min\left\{\frac{m_\lambda}{n_\lambda} + \sqrt{\frac{m_\lambda^2}{n_\lambda^2} - \frac{m_\lambda^2\left[\left(1 - \frac{t_{n-1;1-\alpha}^2}{m_\lambda}\right)n - 1 + t_{n-1;1-\alpha}^2\right]}{n_\lambda^2\left[\left(1 - \frac{t_{n-1;1-\alpha}^2}{n_\lambda}\right)n - 1 + t_{n-1;1-\alpha}^2\right]}}, 1\right\}.$$

The range of possible values for the conditional probability is almost identical to the range of possible values for the unconditional probability if $m_\lambda = m$ and $n_\lambda = n$. The difference becomes smaller as $n_\lambda$ increases and $n$ decreases.

**Example.** In Fig. 5, the possible results for $p^u$ and $p^o$ are shown as an example for $n = 60$ and $0 < m < 60$. The possible values for $p^u$ lie in the lower marked area and the possible values for $p^o$ lie in the upper marked area. The circles represent the results for $\mu_C(Y(k)) \in \{0;1\} \wedge \mu_S(X(k)) \in \{0;1\}$.

For a more detailed view, a section of the upper area is presented in Fig. 6. The range of values of $m$ is $38 < m < 41$. The large circles are adopted from Fig. 5. The dotted line is the upper limit of $p^o$. The stars

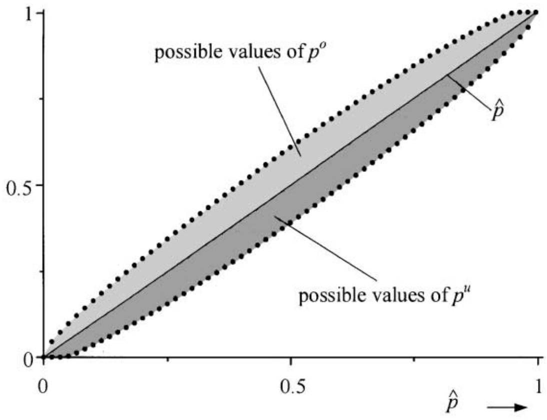



Fig. 5. Using the asymptotic fuzzy relevance test, the possible values for the confidence limits lie in the marked area ($n = 60$ and $0 < m < 60$). The exact values of realization depend on the estimated variances. The circles represent the results for crisp degrees of membership.

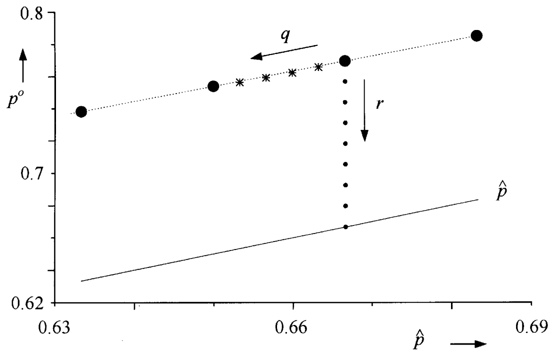



Fig. 6. A section of the upper area of Fig. 5 is presented ($n = 60$ and $38 < m < 41$). Exemplary values of the confidence limit $p^o$ are calculated by the asymptotic fuzzy relevance test: The small circles show how the confidence limit $p^o$ gets smaller, if the degrees of membership are assimilated to each other until all degrees of membership have the same value of 2/3. The stars show that the confidence limit $p^o$ moves along the border of the upper area if only one degree of membership is not crisp and varies from 1 to 0.

represent the results for the following values of $\mu_C(y(k))$:

$$\begin{aligned} \mu_C(y(1)) &= 0, \\ \mu_C(y(2)) &= 0, \\ &\vdots \\ \mu_C(y(20)) &= 0, \\ \mu_C(y(21)) &= 1, \\ \mu_C(y(22)) &= 1, \\ &\vdots \\ \mu_C(y(59)) &= 1, \\ \mu_C(y(60)) &= 1 - q/5 \end{aligned}$$

with $q = 1, 2, 3, 4$. The data samples are constructed in such a way that the value of $\hat{p}$ is decreased iteratively from 40/60 to 39/60 by changing only one data sample.

The small circles represent the results for the following values of $\mu_C(y(k))$:

$$\begin{aligned} \mu_C(y(1)) &= 0 + r/12, \\ \mu_C(y(2)) &= 0 + r/12, \\ &\vdots \\ \mu_C(y(20)) &= 0 + r/12, \end{aligned}$$

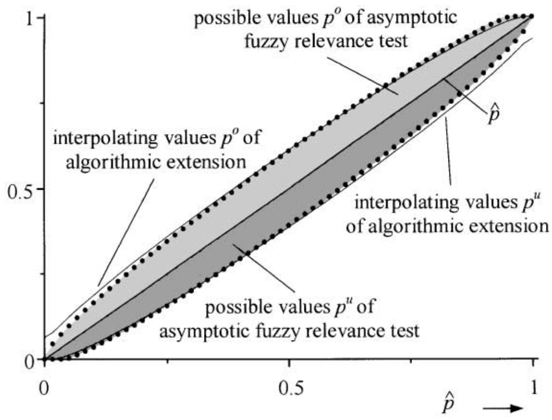



Fig. 7. Comparison of the confidence limits $p^u$ and $p^o$ of the asymptotic fuzzy relevance test (represented by the marked area) and the algorithmic extension (represented by the thin lines) for $n = 60$ and $0 < m < 60$.

$$\mu_C(y(21)) = 1 - r/24,$$
$$\mu_C(y(22)) = 1 - r/24,$$
$$\vdots$$
$$\mu_C(y(60)) = 1 - r/24$$

with $r = 1, 2, \ldots, 8$. The data samples are constructed in such a way that $\hat{p}$ remains constant $\frac{40}{60}$ while all data samples of zero (one) are simultaneously increased (decreased) until all data samples have the same value of $\frac{2}{3}$. For the conditional probability an equivalent diagram to Fig. 6 can be constructed.

Usually, the range of values leading to degrees of membership greater than zero is only a part of the whole range of values covered by data samples. Then, the confidence limits lie mainly near the maximum values.

As the calculations are asymptotic, problems arise for smaller numbers of data samples and here, especially, for the calculation of $p_\lambda^u$ if $\hat{p}_\lambda \approx 1$ (positive rule) and for the calculation of $p_\lambda^o$ if $\hat{p}_\lambda \approx 0$ (negative rule). This results from the fact that for $\hat{p}_\lambda = 0$ there is $p_\lambda^o = 0$ and for $\hat{p}_\lambda = 1$ there is $p_\lambda^u = 1$ independent of the number of data samples. Consequently, rules that are correct for almost all data samples will be seen as relevant even if the number of data samples is small.

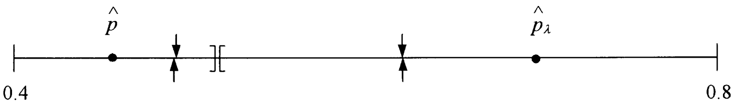
## 5. Comparison

In this section, first, the results of the algorithmic extension of the crisp relevance test are compared with the results of the asymptotic fuzzy relevance test. Afterwards, a concrete set of data samples from a chemical reactor is used to compare all three approaches by means of three examples of IF/THEN statements.

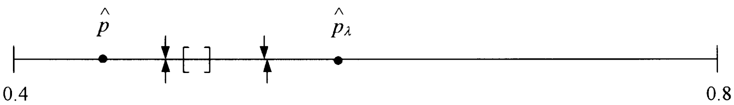
In Fig. 7 the interpolating values of $p^o$ and $p^u$ of the algorithmic extension of the crisp relevance test are represented, together with the possible values of $p^o$ and $p^u$ of the asymptotic fuzzy relevance test of Fig. 5 ($n = 60$ and $0 \leqslant m \leqslant 60$). The interpolating values of the crisp relevance test lie near the lower and upper limit of possible values of the asymptotic fuzzy relevance test. A further comparison is interesting with respect to the following two viewpoints:

- How good is the result of the asymptotic fuzzy relevance test in the special case of crisp sets ($\mu_C(Y(k)) \in \{0; 1\} \wedge \mu_S(X(k)) \in \{0; 1\}$)?


(a) IF (oil temperature is in set 1) THEN (reactor temperature is in set 4) ($n = 206$, $m = 93.86$, $n_\lambda = 15.51$, $m_\lambda = 0.13$).




(b) IF (oil temperature is in set 4) THEN (reactor temperature is in set 4) ($n = 206$, $m = 93.86$, $n_\lambda = 25.69$, $m_\lambda = 17.91$).




(c) IF (oil temperature is in set 2) THEN (reactor temperature is in set 5) ($n = 206$, $m = 92.85$, $n_\lambda = 98.95$, $m_\lambda = 57.76$).

Fig. 8. Comparison of the confidence limits for process data of 206 data samples for three selected IF/THEN statements: ']', '[' results of the algorithmic extension, '↑' results of the Bootstrap fuzzy relevance test, '↓' results of the asymptotic fuzzy relevance test. It can be seen that the results of the asymptotic fuzzy relevance test are very good as the results of the Bootstrap fuzzy relevance test are always nearby. The confidence limits of the algorithmic extension of the crisp relevance test are more conservative.

The crisp relevance test supplies the exact results. The results of the asymptotic fuzzy relevance test are given by the circles of Fig. 7. The difference between the results is the error of the asymptotic fuzzy relevance test in the case of crisp sets. The error decreases monotonically to zero for increasing *n*.

- How good is the interpolation by the algorithmic extension of the crisp relevance test in the case of fuzzy sets ($\mu_C(Y(k)) \in [0;1] \wedge \mu_S(X(k)) \in [0;1]$)?
  The error is small if only a few of the data samples occur in the range of the fuzzy set of the input situation and in the range of the fuzzy set of the output event. Then, the variances are high and the confidence limits are near the lower and upper limits of the possible values. If the data samples occur mainly at the increasing or decreasing edge of the fuzzy set of the input situation and the fuzzy set of the output event, the variances are low and the error is greater. Principally, trapezoidal fuzzy sets will lead to smaller errors than triangular fuzzy sets, and fuzzy sets with low density to greater errors than fuzzy sets with high density.

Using the algorithmic extension of the crisp relevance test, the confidence intervals are mostly greater than necessary. This can be interpreted as a conservative relevance test that corresponds to a minimum confidence coefficient of $1-\alpha$. Consequently, it can happen that statements are not accepted as rules that would be accepted if an exact confidence coefficient of $1-\alpha$ is used.

In accordance with the more complex formula, the computing time of the asymptotic fuzzy relevance test is a little longer than the computing time of the crisp relevance test, whereas the computing time of the Bootstrap fuzzy relevance test is not practical for testing a higher number of statements. Nevertheless, the Bootstrap fuzzy relevance test can be used to judge the results of the other two relevance tests, as it supplies very good results [5,17].

In Fig. 8, the results of the three relevance tests of three different statements are shown. The confidence intervals are calculated with a confidence coefficient of 0.95. Measured by the results of the Bootstrap fuzzy relevance test, the results of the asymptotic fuzzy relevance test are very good. The confidence intervals of the algorithmic extension of the crisp relevance test are larger. The first two statements are seen as relevant by all three tests as the confidence intervals do not overlap. The first statement represents a negative relevant rule, the second statement a positive relevant rule. The third statement is seen as a positive relevant rule by the fuzzy relevance tests, but not by the algorithmic extension of the crisp relevance test. Here, the larger confidence intervals cause an overlap.

## 6. Conclusions

In the field of data-based fuzzy modelling, the incremental accumulation of single relevant rules allows complex problems to be handled. To decide if an IF/THEN statement is a relevant rule a relevance test is necessary. A statistical approach is given by the demand that the confidence intervals of the probabilities $p$ and $p_\lambda$ do not overlap – with $p$ being the probability that the output event of the conclusion is true and $p_\lambda$ being the probability that the output event is true under the condition that the input situation of the premise is true.

For crisp rule-based modelling the confidence intervals can be calculated by conventional statistical formulae. For fuzzy modelling, problems arise. Three different solutions are proposed in this paper: an algorithmic extension of the crisp relevance test, a Bootstrap fuzzy relevance test, and an asymptotic fuzzy relevance test.

The algorithmic extension is the quickest relevance test. It is best if many statements are tested and if the conservativeness of the relevance test is not disadvantageous. This is the case if there is a multitude of relevant redundant rules in the search space. In the other cases, the higher calculation effort of the fuzzy relevance tests will achieve the desired result. The asymptotic fuzzy relevance test is appropriate for more than 40 data samples, and the Bootstrap fuzzy relevance test for few data samples and a small number of statements to be tested.

The concept of the relevance test is successfully used in several industrial applications in the context of the Fuzzy-ROSA method [12,15,19].

## Acknowledgements

We thank Prof. Dr. Hering and Dr. Poehlmann for constructive statistical discussions. The research is sponsored by the Deutsche Forschungsgemeinschaft (DFG), as part of the Collaborative Research Center 'Computational Intelligence' (531) of the University of Dortmund.

## References

[1] H. Bandemer, S. Gottwald, Einführung in Fuzzy-Methoden: Theorie und Anwendung unscharfer Mengen, Akademie GmbH, Berlin, 1993.
[2] D. Dubois, H. Prade, Fuzzy Sets and Systems: Theory and Applications, Academic Press, New York, USA, 1980.
[3] B. Efron, Bootstrap methods: another look at the jackknife, Ann. Statist. 7 (1979) 1–26.
[4] B. Efron, Better bootstrap confidence intervals, J. Amer. Statist. Assoc. 82 (1987) 171–200.
[5] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, USA, 1993.
[6] E.C. Fieller, Some problems in interval estimation, J. Roy. Statist. Soc. Ser. B16 (1954) 175–185.
[7] J. Hartung, B. Elpelt, K.-H. Kloesener, Statistik, Oldenbourg, München, 1995.
[8] J.S.U. Hjorth, Computer Intensive Statistical Methods, Chapman & Hall, New York, USA, 1994.
[9] H. Kiendl, R. Knicker, F. Niewels, Two way fuzzy controllers based on hyperinference and inference filter, Proc. World Automation Congress, WAC '96, Montpellier, Frankreich, 1996.

[10] M. Krabs, H. Kiendl, Anwendungsfelder der automatischen Regelgenerierung mit dem ROSA-Verfahren, Automatisierungstechnik 43 (6) (1995) 269–276.
[11] A. Krone, Advanced rule reduction concepts for optimising efficiency of knowledge extraction, Proc. 4th European Congress on Intelligent Techniques and Soft Computing, EUFIT '96, vol. 2, Aachen, 1996, pp. 919–923.
[12] A. Krone, C. Frenck, O. Russak, Design of a fuzzy controller for an alkoxylation process using the ROSA method for automatic rule generation, Proc. 3rd European Congress on Intelligent Techniques and Soft Computing, EUFIT '95, vol. 2, Aachen, 1995, pp. 760–764.
[13] A. Krone, H. Kiendl, Rule-based decision analysis with Fuzzy-ROSA method, Proc. European Workshop on Fuzzy Decision Analysis for Management, Planning and Optimization, EFDAN '96, Dortmund, 1996, pp. 109–114.
[14] A. Krone, H. Kiendl, Evolutionary concept for generating relevant fuzzy rules from data, Internat. J. Knowledge-based Intelligent Eng. Systems 1 (4) (1997) 207–213.
[15] A. Krone, U. Schwane, Generating fuzzy rules from contradictory data of different control strategies and control performances, Proc. 5th IEEE Internat. Conf. on Fuzzy Systems, FUZZ-IEEE '96, vol. 1, New Orleans, USA, 1996, pp. 492–497.
[16] M. Mizumoto, Pictorial representations of fuzzy connectives, part I: Cases of t-norms, t-conorms and averaging operators, Fuzzy Sets and Systems 31 (1989) 217–242.
[17] J. Shao, D. Tu, The Jackknife and Bootstrap, Springer, New York, USA, 1995.
[18] T. Slawinski, A. Krone, U. Hammel, D. Wiesmann, P. Krause, A hybrid evolutionary search concept for data-based generation of relevant fuzzy rules in high dimensional spaces, Proc. 8th Internat. Conf. on Fuzzy Systems, FUZZ-IEEE'99, Korea, 1999.
[19] T. Slawinski, J. Praczyk, U. Schwane, A. Krone, H. Kiendl, Data-based generation of fuzzy-rules for classification, prediction and control with the Fuzzy-ROSA method, European Control Congress, ECC '99, Karlsruhe, 1999.
[20] P. Smets, Probability of a fuzzy event: an axiomatic approach, Fuzzy Sets and Systems 7 (1982) 153–164.
[21] H. Taeger, Relevanzüberprüfung von unscharfen Regeln mittels statistischer Verfahren, Master's Thesis, University of Dortmund, Faculty of Statistics, 1998.
[22] L.A. Zadeh, Fuzzy sets, Inform. Control 8 (1965) 338–353.
[23] L.H. Zadeh, Probability measures of fuzzy events, J. Math. Anal. Appl. 23 (1968) 421–427.